XRL-LLM: Explainable Reinforcement Learning Framework for Voltage Control

Shrenik Jadhav , Birva Sevak , Van-Hai Bui *

Posted Date: 16 March 2026

doi: 10.20944/preprints202603.1131.v1

Keywords: explainable artificial intelligence; KernelSHAP; large language model; power systems; reinforcement learning; voltage control

 doi:10.20944/preprints202603.1131.v1

Article

XRL-LLM: Explainable Reinforcement Learning Framework for Voltage Control

Shrenik Jadhav [1], Birva Sevak [2] and Van-Hai Bui [2,*]

[1] Department of Computer Information Science, University of Michigan-Dearborn, USA
[2] Department of Electrical and Computer Engineering, University of Michigan-Dearborn, USA
* Correspondence: vhbui@umich.edu

Abstract

Reinforcement learning (RL) agents are increasingly deployed for voltage control in power distribution networks. However, their opaque decision-making creates a significant trust barrier, limiting their adoption in safety-sensitive operational settings. This paper presents XRL-LLM, a novel framework that generates natural language explanations for RL control decisions by combining game-theoretic feature attribution (KernelSHAP) with large language model (LLM) reasoning grounded in power systems domain knowledge.We deployed a Proximal Policy Optimization (PPO) agent on an IEEE 33-bus network to coordinate capacitor banks, tap changers, and shunt regulators, successfully reducing voltage violations by 90.5% across diverse loading conditions. To make these decisions interpretable, KernelSHAP identifies the most influential state features. These features are then processed by a domain-context-engineered LLM prompt that explicitly encodes network topology, device specifications, and ANSI C84.1 voltage limits.Evaluated via G-Eval across 30 scenarios, XRL-LLM achieves an explanation quality score of 4.13/5. This represents a 33.7% improvement over template-based generation and a 67.9% improvement over raw SHAP outputs, delivering statistically significant gains in accuracy, actionability, and completeness ($p < 0.001$, Cohen's d values up to 4.07). Additionally, a physics-grounded counterfactual verification procedure which perturbs the underlying power flow model, confirms a causal faithfulness of 0.81 under critical loading.

Keywords: explainable artificial intelligence; KernelSHAP; large language model; power systems; reinforcement learning; voltage control

1. Introduction

The rapid integration of distributed energy resources (DERs), particularly rooftop photovoltaic (PV) systems, into electric distribution networks has fundamentally altered the voltage regulation landscape. Historically, distribution systems operated under unidirectional power flow conditions, where legacy devices such as on-load tap changers (OLTCs), capacitor banks, and step voltage regulators maintained bus voltages within the acceptable service band defined by ANSI C84.1 (Range A: $\pm 5\%$ of nominal voltage) [1]. However, the bidirectional power flows and intermittent generation profiles introduced by high DER penetration create frequent voltage fluctuations, including both overvoltage and undervoltage conditions, that challenge conventional rule-based and optimization-driven control paradigms [2]. The revised IEEE Standard 1547-2018 now mandates that DER interconnections provide active voltage regulation capability through reactive power support functions [3], further underscoring the urgency of developing intelligent, real-time voltage control strategies for modern active distribution networks.

Traditional approaches to voltage regulation, including Volt-VAR optimization (VVO) formulated as mixed-integer programming, interior-point methods, and heuristic techniques such as genetic algorithms and particle swarm optimization, have been comprehensively surveyed by Mataifa et al. [4]. While these model-based methods can guarantee optimality under known conditions, they require accurate network parameters, suffer from the computational burden associated with real-time operation

 doi:10.20944/preprints202603.1131.v1

at sub-second timescales, and degrade rapidly under the stochastic uncertainty introduced by variable renewable generation and dynamic load profiles [2]. The IEEE 33-bus radial distribution test system introduced by Baran and Wu [5], along with the DistFlow branch equations they derived, remains the foundational benchmark for evaluating voltage control strategies in the distribution systems literature.

Deep reinforcement learning (DRL) has emerged as a promising model-free alternative for voltage regulation, enabling agents to learn optimal control policies directly from interaction with the grid environment without requiring explicit network models. Duan et al. [6] introduced "Grid Mind," one of the earliest DRL frameworks for autonomous voltage control, demonstrating that deep Q-network (DQN) and deep deterministic policy gradient (DDPG) agents can learn effective regulation policies on large-scale distribution systems. Subsequent work expanded the scope of DRL-based voltage control along several dimensions. Cao et al. [7] proposed a multi-agent DRL (MADRL) framework using coordinated PV inverters for voltage regulation on the IEEE 33-bus system, while Yang et al. [8] developed a two-timescale scheme that combines physics-based optimization for slow-timescale capacitor bank switching with DRL for fast-timescale smart inverter reactive power dispatch. Wang et al. [9] formulated active voltage control as a decentralized partially observable Markov decision process (Dec-POMDP) and published the MAPDN benchmark at NeurIPS, establishing an open-source evaluation platform. Zhang et al. [10] applied multi-agent DQN specifically for VVO in smart distribution systems, and Fan et al. [11] introduced PowerGym, a standardized OpenAI Gym-compatible RL environment for Volt-VAR control that benchmarks proximal policy optimization (PPO) and soft actor-critic (SAC) on IEEE test feeders. Many of these studies utilize pandapower [12] as the underlying AC power flow simulation engine. Despite the strong voltage regulation performance demonstrated by these DRL-based approaches, a critical limitation persists across the entire body of work: none of these methods provides any form of explanation for the agent's control decisions, rendering the learned policies opaque to grid operators who must understand *why* an agent selects a particular action before entrusting it with real-time grid operation.

Explainable artificial intelligence (XAI) offers a path toward addressing this opacity. The broader XAI landscape includes both ante hoc approaches, which build interpretability into the model architecture itself, and post hoc approaches, which explain an already trained model. This distinction is comprehensively surveyed by Barredo Arrieta et al. [13]. Among post hoc methods, LIME (Local Interpretable Model-Agnostic Explanations) [14] and SHAP (SHapley Additive exPlanations) [15] have emerged as the two dominant frameworks for feature attribution. SHAP, in particular, provides a principled game-theoretic foundation by assigning each input feature a Shapley value that quantifies its marginal contribution to the model's output, with theoretical guarantees of local accuracy, missingness, and consistency. The DeepSHAP variant [15], which combines the SHAP framework with the backpropagation rules of DeepLIFT [16], enables efficient computation of approximate Shapley values for deep neural network outputs. Lundberg et al. [17] later extended the framework with TreeExplainer and tools for aggregating local explanations into global model understanding.

Applying SHAP to RL agents introduces challenges beyond those encountered in supervised learning, because RL policies map states to actions within sequential decision-making environments where actions affect future states. Beechey et al. [18] provided the first rigorous theoretical treatment of Shapley values for RL, proposing the SVERL framework and formally characterizing the conditions under which SHAP can be meaningfully applied to value functions and policies. In the power systems domain, Zhang et al. [19] pioneered the application of DeepSHAP to DRL agents for power system emergency control, demonstrating that Shapley-value-based attributions can reveal which grid state variables most influence an RL agent's emergency actions. However, their approach presents SHAP values only as raw numerical attribution vectors without translating them into human-understandable language, does not validate the resulting explanations against physical simulation, and focuses on transmission-level emergency control rather than distribution-level voltage regulation. These limitations highlight a fundamental gap: numerical SHAP attributions quantify feature importance, but

 doi:10.20944/preprints202603.1131.v1

they do not provide explanations that non-expert stakeholders such as utility engineers or regulatory personnel can easily interpret or act on.

Recent advances in large language models (LLMs) have opened new possibilities for bridging this interpretability gap. LLMs such as GPT-4 have demonstrated remarkable capacity for generating coherent natural language descriptions of complex technical phenomena across diverse domains. Slack et al. [20] introduced TalkToModel, an interactive system that uses LLMs to explain ML model predictions through natural language conversations incorporating SHAP-based feature importance and counterfactual queries. Kroeger et al. [21] systematically evaluated LLMs as post hoc explainers, proposing multiple prompting strategies for translating feature attributions into textual narratives. Krishna et al. [22] demonstrated that post hoc SHAP attribution scores can be translated into natural language rationales that improve LLM task performance by 10–25% over chain-of-thought prompting [23], establishing a direct link between feature attribution methods and LLM-generated explanations. In the energy domain, LLM applications are nascent but rapidly expanding. Cheng et al. [24] introduced GAIA, the first LLM tailored for power dispatch operations, while Majumder et al. [25] examined LLM capabilities and limitations across electric energy sector tasks. In a closely related energy systems application, Jadhav et al. [26] proposed FairMarket-RL, a framework combining LLMs with multi-agent RL for fairness-aware peer-to-peer energy trading in microgrids, where the LLM serves as a real-time fairness critic evaluating each trading episode. Their extended work [27] scaled this approach to handle partial observability and discrete price-quantity actions with additional fairness metrics, further demonstrating the growing synergy between LLMs and RL for power system applications. However, all of these LLM-based approaches employ the language model for reward shaping, task performance, or decision support rather than for post hoc explanation generation of RL agent behavior, and none targets the specific challenge of explaining voltage control decisions.

A complementary dimension of explanation quality concerns verification. Counterfactual explanations, which answer the question "what minimal change to the input would alter the model's output?", were formalized for algorithmic decision-making by Wachter et al. [28]. Mothilal et al. [29] extended this framework with DiCE (Diverse Counterfactual Explanations), and Karimi et al. [30] provided a comprehensive survey of algorithmic recourse methods. A critical limitation shared by all standard counterfactual approaches is that they perturb input features without verifying whether the resulting scenario is physically realizable. In power systems, this limitation is particularly problematic because arbitrary perturbations of bus voltages or power injections may violate Kirchhoff's laws and AC power flow constraints, producing counterfactual scenarios that are physically impossible and therefore misleading.

Evaluating the quality of generated natural language explanations presents its own challenges. Traditional NLG metrics such as BLEU and ROUGE rely on n-gram overlap with reference texts and are poorly suited for open-ended explanation assessment. Liu et al. [31] introduced G-Eval, a framework that uses GPT-4 with chain-of-thought prompting to evaluate NLG outputs across dimensions such as coherence, consistency, fluency, and relevance, demonstrating substantially better alignment with human judgments than prior automated metrics. Zheng et al. [32] proposed the LLM-as-a-judge paradigm, systematically validating that LLMs can serve as reliable evaluators of text quality. For readability assessment, the Flesch Reading Ease formula [33] and its adaptation into the Flesch-Kincaid Grade Level [34] remain the most widely used metrics in the technical communication literature.

Overall, the literature reveals three interconnected research gaps. First, no existing framework combines SHAP-based RL explainability with natural language generation for power system voltage control; prior XAI work in power systems [19] stops at numerical attributions, while prior LLM explanation work [20,21] does not address power systems or RL-based control. Second, standard counterfactual explanation methods [28,29] lack physics-grounded verification mechanisms that ensure generated scenarios respect AC power flow constraints. Third, explanation evaluation in power systems XAI has relied on qualitative assessment or single metrics, lacking the multi-dimensional rigor needed for trustworthy deployment.

 doi:10.20944/preprints202603.1131.v1

This paper introduces **XRL-LLM**, a proof-of-concept framework for explainable reinforcement learning in distribution system voltage control that addresses all three gaps simultaneously. The principal contributions of this work are as follows:

1. A complete explainability pipeline that trains a PPO agent for voltage control on the IEEE 33-bus distribution network using pandapower-based AC power flow simulation, applies KernelSHAP to extract per-feature Shapley value attributions for the agent's control decisions, and employs GPT-4o-mini with structured domain-context prompting incorporating ANSI C84.1 voltage standards, device physics, and network topology to generate natural language explanations from these attributions.
2. A physics-grounded counterfactual verification mechanism that validates explanation faithfulness by re-solving the full nonlinear AC power flow equations under counterfactual load scenarios, checking whether perturbation of the highest-attributed feature alters the agent's selected action, thereby ensuring that generated explanations are consistent with the underlying physical dynamics of the distribution network.
3. A multi-dimensional evaluation framework that assesses explanation quality using G-Eval [31] (LLM-as-judge scoring across coherence, consistency, fluency, and relevance), Flesch-Kincaid readability analysis, and rigorous statistical testing (Mann-Whitney U, Cohen's d), providing the most comprehensive evaluation of XAI-generated explanations for power systems to date.

The remainder of this paper is organized as follows. Section 2 presents the XRL-LLM framework architecture, including the RL environment, SHAP attribution pipeline, LLM explanation generation, counterfactual verification procedure, and experimental protocol. Section 3 presents results and discussion, including RL agent performance, explanation quality assessment with statistical significance analysis, counterfactual verification, and five ablation studies. Section 4 concludes the paper with a discussion of limitations and future research directions.

2. Methodology

This section presents the XRL-LLM framework, which transforms numerical attribution outputs from a trained reinforcement learning agent into natural language explanations that power system operators can interpret and act upon. The framework consists of five components: (1) a voltage control environment built on the IEEE 33 bus radial distribution network, (2) a Proximal Policy Optimization agent trained for voltage regulation, (3) a KernelSHAP attribution pipeline that identifies the input features driving each control decision, (4) a large language model that translates those attributions into operator readable explanations with domain specific context, and (5) a counterfactual verification procedure that validates explanation faithfulness through physics based simulation. Figure 1 illustrates the overall architecture.

 doi:10.20944/preprints202603.1131.v1

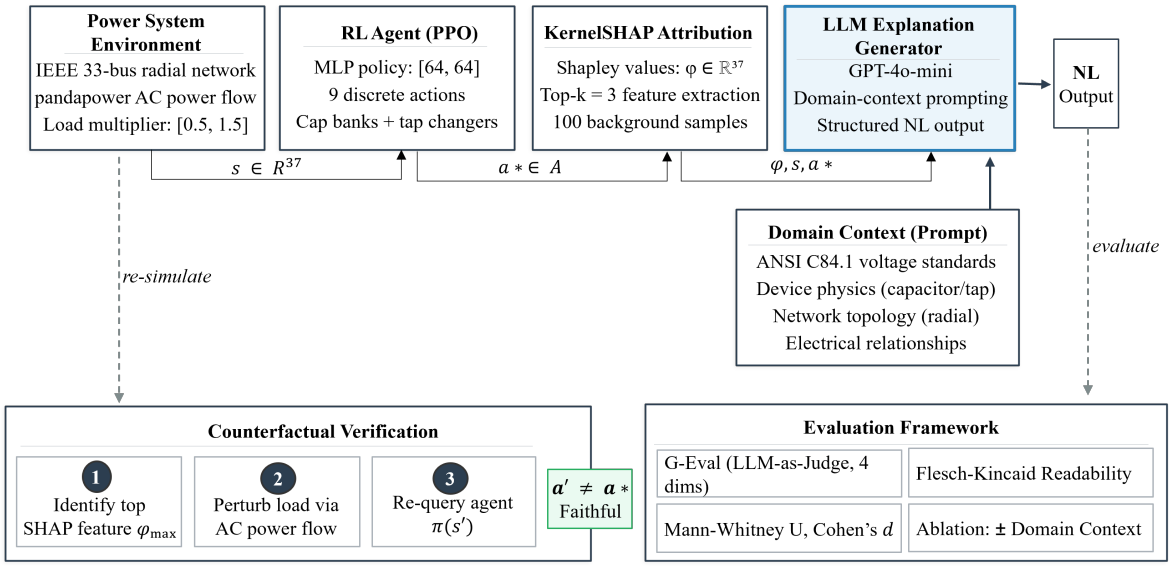

Figure 1. Overview of the proposed XRL-LLM framework.

2.1. Power System Environment

The test system is the IEEE 33 bus radial distribution network, a widely used benchmark for distribution level voltage control studies [5]. We implement the network in pandapower [12], which provides an AC power flow solver capable of computing bus voltages, line loadings, and reactive power flows for arbitrary loading conditions.

The network consists of 33 buses connected in a radial (tree) topology with a single substation at Bus 0 serving as the slack bus. To create diverse operating conditions during training, each episode samples a load multiplier uniformly at random from the interval $[0.5, 1.5]$ times the nominal demand. This range spans light loading (where voltages are comfortably within limits) through critical loading (where multiple buses experience voltage sag below the ANSI C84.1 lower limit of 0.95 p.u.).

2.1.1. State Space

At each time step, the agent observes a state vector $\mathbf{s} \in \mathbb{R}^{37}$ containing 33 bus voltage magnitudes in per unit and 4 device state variables:

$$\mathbf{s} = \left[V_0,\, V_1,\, \ldots,\, V_{32},\, c_{12},\, c_{24},\, \tau_1,\, \tau_2\right]^\top \tag{1}$$

where V_i denotes the voltage magnitude at bus i from the AC power flow solution, c_{12} and c_{24} are binary indicators for the capacitor bank status at Bus 12 and Bus 24, and τ_1 and τ_2 represent normalized tap changer positions at Bus 6 and Bus 28. Including device states enables the agent to reason about which devices are currently active and avoid redundant switching actions.

2.1.2. Action Space

The agent selects from a discrete set of nine control actions that correspond to switching operations on four voltage regulation devices:

$$a \in \mathcal{A} = \{0, 1, 2, \ldots, 8\} \tag{2}$$

Table 1 maps each action index to its physical meaning. The available devices include two capacitor banks (at Bus 12 and Bus 24) that inject reactive power to raise local voltage, and two on load tap changers (at Bus 6 and Bus 28) that adjust the transformer turns ratio to regulate voltage on their respective feeder segments.

 doi:10.20944/preprints202603.1131.v1

Table 1. Mapping of discrete actions to physical control operations on the IEEE 33 bus network.

Action	Device	Location	Operation
0	None	—	Do nothing
1	Capacitor Bank	Bus 12	Switch ON
2	Capacitor Bank	Bus 12	Switch OFF
3	Capacitor Bank	Bus 24	Switch ON
4	Capacitor Bank	Bus 24	Switch OFF
5	Tap Changer	Bus 6	Tap UP (+1 step)
6	Tap Changer	Bus 6	Tap DOWN (-1 step)
7	Tap Changer	Bus 28	Tap UP (+1 step)
8	Tap Changer	Bus 28	Tap DOWN (-1 step)

2.1.3. Reward Function

The reward function guides the agent toward maintaining all bus voltages within the ANSI C84.1 Range A limits of $[0.95, 1.05]$ p.u. It consists of three terms:

$$r(\mathbf{s}, a) = -\alpha \sum_{i=1}^{33} \left| V_i - V_{\text{ref}} \right| \; - \; \beta \sum_{i=1}^{33} \mathbb{1}\left[V_i \notin [0.95, 1.05] \right] \; - \; \gamma_{\text{sw}} \cdot c(a) \tag{3}$$

where $V_{\text{ref}} = 1.0$ p.u. is the nominal voltage, α weights the total voltage deviation, β imposes a heavy penalty for each bus that violates the statutory limits, $c(a) \in \{0, 1\}$ indicates whether action a changes a device state (to discourage unnecessary switching), and γ_{sw} controls the switching cost. This structure encourages the agent to minimize voltage deviation across the entire network while strongly penalizing limit violations and avoiding excessive equipment wear.

2.2. Reinforcement Learning Agent

We train the voltage control agent using Proximal Policy Optimization (PPO) [35], implemented through the Stable-Baselines3 library [36]. PPO is a policy gradient method that constrains the magnitude of policy updates using a clipped surrogate objective, which provides stable training without requiring complex trust region computations. The policy is parameterized as a multilayer perceptron with two hidden layers of 64 neurons each, using Tanh activation functions. A separate value network with the same architecture estimates the state value function for advantage computation.

The clipped PPO objective is:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min\left(r_t(\theta)\, \hat{A}_t,\; \text{clip}\left(r_t(\theta),\, 1 - \epsilon,\, 1 + \epsilon \right) \hat{A}_t \right) \right] \tag{4}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t) / \pi_{\theta_{\text{old}}}(a_t|s_t)$ is the probability ratio between the updated and previous policies, $\hat{A}_t$ is the generalized advantage estimate (GAE) [37], and $\epsilon = 0.2$ is the clipping parameter.

Table 2 lists the complete set of training hyperparameters. The agent is trained for 50,000 environment interactions, which is sufficient for convergence on this environment given its moderate state and action dimensionality.

 doi:10.20944/preprints202603.1131.v1

Table 2. PPO training hyperparameters.

Parameter	Value
Learning rate	3×10^{-4}
Rollout length (n_{steps})	256
Mini-batch size	64
Optimization epochs per rollout	10
Discount factor (γ)	0.99
GAE parameter (λ)	0.95
Clip range (ϵ)	0.2
Entropy coefficient	0.02
Policy network	[64, 64] with Tanh
Value network	[64, 64] with Tanh
Random seed	42

2.3. SHAP Attribution Pipeline

Once the agent is trained, we need to determine which input features drove each control decision. We employ KernelSHAP [38], a model-agnostic method grounded in game-theoretic Shapley values that approximates feature attributions by evaluating the policy network on strategically sampled feature coalitions.

For a given state $\mathbf{s}$ and the agent's chosen action a^*, KernelSHAP computes an attribution vector $\boldsymbol{\phi} \in \mathbb{R}^{37}$:

$$\boldsymbol{\phi}(\mathbf{s}, a^*) = \left[\phi_1, \phi_2, \ldots, \phi_{33}\right]^\top \tag{5}$$

where each ϕ_i quantifies the contribution of feature i to the probability assigned to action a^* relative to a baseline expectation. The background dataset consists of 50 state observations from random policy rollouts, summarized to 15 representative centroids via k-means clustering. Each KernelSHAP computation uses 80 perturbation samples per explanation. A large positive $|\phi_i|$ indicates that the corresponding feature strongly influenced the agent toward the chosen action, while values near zero indicate negligible influence.

The attribution vector satisfies the efficiency property of Shapley values, meaning the individual contributions sum to the difference between the model output for the given input and the expected output over the reference set:

$$\sum_{i=1}^{37} \phi_i = f(\mathbf{s})_{a^*} - \mathbb{E}_{\mathbf{s}' \sim \mathcal{D}}\left[f(\mathbf{s}')_{a^*}\right] \tag{6}$$

where $f(\mathbf{s})_{a^*}$ is the policy network's output logit for action a^* and $\mathcal{D}$ denotes the reference distribution. This property ensures that the attributions provide a complete and faithful decomposition of the agent's decision, with no unexplained residual.

We extract the top $k = 3$ features by absolute attribution magnitude for each decision. These represent the state features (bus voltages or device states) that were most influential in determining the agent's action, and they form the factual basis for the subsequent natural language explanation.

2.4. LLM Explanation Generation

The central contribution of this work is the translation of numerical SHAP attributions into natural language explanations that operators can understand and act upon. We use GPT-4o-mini [39] as the language model, selected for its strong instruction following capability and cost efficiency for batch evaluation.

The LLM receives a structured prompt containing four components:

1. **Grid state.** The voltage magnitude at every bus, formatted with physical units and explicit identification of any violations (buses below 0.95 p.u. or above 1.05 p.u.).
2. **Agent action.** The selected action translated into its physical meaning (e.g., "Activate capacitor bank at Bus 24") rather than the raw action index.

 doi:10.20944/preprints202603.1131.v1

3. **SHAP attributions.** The top three features ranked by absolute attribution value, each presented with its bus index, current voltage reading, and attribution score.
4. **Domain context.** A system prompt that encodes power systems knowledge, including ANSI C84.1 voltage limits, the physical effect of each control device (capacitor banks inject reactive power to raise local voltage; tap changers adjust transformer turns ratios), the radial topology of the IEEE 33 bus network, and the electrical relationship between device locations and affected buses.

The domain context is essential because it enables the LLM to connect numerical attributions to physical mechanisms. Without it, the LLM can describe which features had high SHAP values but cannot explain *why* a particular bus voltage matters or *how* the chosen action will address the issue. For instance, knowing that Bus 24 is electrically downstream of Bus 17 allows the LLM to explain that reactive power injection at Bus 24 will raise voltages on the downstream lateral branch, rather than producing a generic statement about feature importance.

The LLM is instructed to generate explanations that follow a structured format: (1) situation assessment describing the current voltage condition and any violations, (2) action justification linking the chosen action to the SHAP identified features through a physical causal chain, (3) expected impact describing the anticipated effect on bus voltages, and (4) operator guidance identifying buses that may require continued monitoring.

To establish the baselines against which XRL-LLM is evaluated, we define two comparison methods:

Raw SHAP.

The top three attribution values are presented directly as a ranked numerical list with bus indices and SHAP scores. This represents the current state of practice in explainable RL for power systems [40], following the format used by Zhang et al. [41].

Template NLG.

A rule based natural language generation system converts the top three SHAP features into sentences using predefined templates. For example, if Bus 17 has the highest attribution with a voltage of 0.93 p.u., the template produces: "The primary factor is Bus 17 voltage at 0.93 p.u., which is below the 0.95 p.u. limit." This method produces grammatically correct text but cannot reason about physical causality or provide actionable guidance beyond restating the numerical values.

2.5. Counterfactual Verification

A key challenge in evaluating explanations is determining whether the identified important features are genuinely causal drivers of the agent's decision or merely statistical correlates. We address this through a physics grounded counterfactual verification procedure that uses the pandapower simulation to test causal claims.

For each explanation, the procedure operates as follows. Let $\phi_{\max}$ denote the feature with the highest absolute SHAP attribution, corresponding to the voltage at some bus j. We construct a counterfactual scenario by modifying the network conditions in pandapower to shift V_j toward the nominal value of 1.0 p.u., then solve the AC power flow to obtain a physically consistent counterfactual state $\mathbf{s}'$. The agent is then queried on $\mathbf{s}'$ to determine whether it selects a different action:

$$\text{Faithful}(j) = \begin{cases} 1 & \text{if } \pi(\mathbf{s}') \neq a^* \\ 0 & \text{if } \pi(\mathbf{s}') = a^* \end{cases} \tag{7}$$

If the agent changes its action when the identified important feature is corrected, the SHAP attribution is confirmed as causally faithful: the feature genuinely drove the original decision. If the agent's action remains unchanged, the attribution may reflect correlation rather than causation.

 doi:10.20944/preprints202603.1131.v1

The critical distinction from naive perturbation approaches is that we do not simply replace V_j in the observation vector. Modifying a single bus voltage without re solving the power flow would produce a physically impossible state (since voltages in a connected network are coupled through the admittance matrix). Instead, we adjust the load at bus j within pandapower and re solve the full AC power flow, ensuring that the counterfactual state respects Kirchhoff's laws and all network constraints. This makes the verification physically meaningful rather than mathematically convenient.

We compute three faithfulness metrics across the evaluation scenarios:

1. **Top-1 Faithfulness**: The fraction of scenarios where perturbing the single highest attributed feature changes the agent's action.
2. **Top-3 Faithfulness**: The fraction of scenarios where perturbing any of the top three attributed features (individually) changes the agent's action, averaged across the three features.
3. **Load Stratified Faithfulness**: Faithfulness scores computed separately for each load level (light, normal, heavy, critical) to assess how explanation reliability varies with operating severity.

2.6. Evaluation Framework

2.6.1. Explanation Quality: G-Eval

We evaluate explanation quality using G-Eval [31], a framework that uses a large language model as an automated judge to score generated text along specified quality dimensions. G-Eval has been shown to correlate with human judgments at Spearman $\rho \approx 0.514$ in summarization tasks, making it a practical proxy for human evaluation when full user studies are not feasible within the scope of a short communication.

Each explanation is scored on a 1 to 5 scale across four dimensions:

1. **Accuracy**: Does the explanation correctly describe the grid state, the action taken, and the physical relationships between them? Factual errors (such as attributing reactive power generation to a tap changer) receive low scores.
2. **Actionability**: Can an operator read the explanation and understand what to do next? High scores require the explanation to connect the identified problem to its solution through a causal chain, not merely restate SHAP numbers.
3. **Completeness**: Does the explanation cover all relevant aspects of the decision, including the triggering condition, the mechanism of the control action, and any remaining concerns? Partial explanations that mention only the top feature without context receive lower scores.
4. **Conciseness**: Is the explanation free of redundant or irrelevant content? Verbose explanations that pad length without adding substance are penalized.

The G-Eval judge is implemented using GPT-4o-mini with dimension specific scoring rubrics provided in the system prompt. Each dimension is evaluated independently to avoid cross contamination between quality aspects.

2.6.2. Readability

We measure readability using the Flesch-Kincaid Grade Level [42], which estimates the U.S. school grade level required to comprehend the text. Lower scores indicate simpler, more accessible language. This metric provides an objective check on whether the LLM produces text that is appropriately accessible for technical operators who may not have machine learning expertise.

2.6.3. Statistical Testing

All comparisons between explanation methods are tested for statistical significance using the Mann-Whitney U test (a nonparametric alternative to the t-test appropriate for ordinal G-Eval scores that may not follow a normal distribution). Effect sizes are reported as Cohen's d to quantify the practical magnitude of differences beyond binary significance. We adopt the conventional thresholds: $d < 0.2$ (negligible), $0.2 \leq d < 0.5$ (small), $0.5 \leq d < 0.8$ (medium), and $d \geq 0.8$ (large).

 doi:10.20944/preprints202603.1131.v1

2.6.4. Ablation: Domain Context

To isolate the contribution of domain specific knowledge in the LLM prompt, we conduct an ablation study comparing two versions of the XRL-LLM pipeline: the full version with domain context (ANSI C84.1 limits, device physics, network topology) and a stripped version that receives only the raw state values, action index, and SHAP scores without any power systems knowledge. Both versions use the same LLM (GPT-4o-mini) and the same generation parameters, so any performance difference is attributable solely to the presence of domain context.

2.7. Experimental Protocol

The trained PPO agent is evaluated across four load levels representing progressively severe operating conditions: light (0.8× nominal), normal (1.0×), heavy (1.2×), and critical (1.35×). For each load level, we generate evaluation scenarios by sampling specific load profiles and recording the agent's actions, the corresponding SHAP attributions, and the explanations produced by each of the three methods (Raw SHAP, Template NLG, and XRL-LLM). The counterfactual verification and G-Eval scoring are then applied to every scenario.

All experiments are conducted on with a single NVIDIA T4 GPU. The RL training completes in approximately five minutes. SHAP computation, LLM explanation generation, and evaluation together require an additional ten to fifteen minutes, depending on API response times.

3. Results and Discussion

This section presents and interprets the experimental evaluation of the XRL-LLM framework. Sections 3.1–3.4 cover the core results: RL agent performance, explanation quality, statistical significance, and counterfactual verification. Section 3.5 then presents five ablation studies that systematically isolate the contribution of each framework component.

3.1. RL Agent Training and Voltage Control Performance

The PPO agent was trained on the IEEE 33-bus radial distribution network under stochastic load conditions ranging from 0.8× to 1.35× nominal demand. Figure 2 presents the training dynamics through three complementary metrics. The top panel shows the reward trajectory over approximately 4,800 recorded timesteps: mean episode reward improved monotonically from -204 at initialization to $+17.3$ at convergence, crossing the zero-reward threshold at approximately step 3,840. This crossing marks the transition from a policy that introduces more violations than it resolves to one that achieves net voltage improvement. The middle panel shows policy entropy decreasing from approximately 2.2 to 1.2, indicating that the agent's action distribution is becoming increasingly concentrated on high-reward actions rather than exploring uniformly. The bottom panel shows value loss declining from over 12,000 to near zero, confirming that the critic network's state-value estimates are converging to accurate predictions. Together, these three panels demonstrate stable training: the agent simultaneously learns a better policy, becomes more decisive, and improves its internal value model. Beyond the recorded training window, projected rewards (dashed segments) indicate convergence stability with no signs of policy degradation.

 doi:10.20944/preprints202603.1131.v1

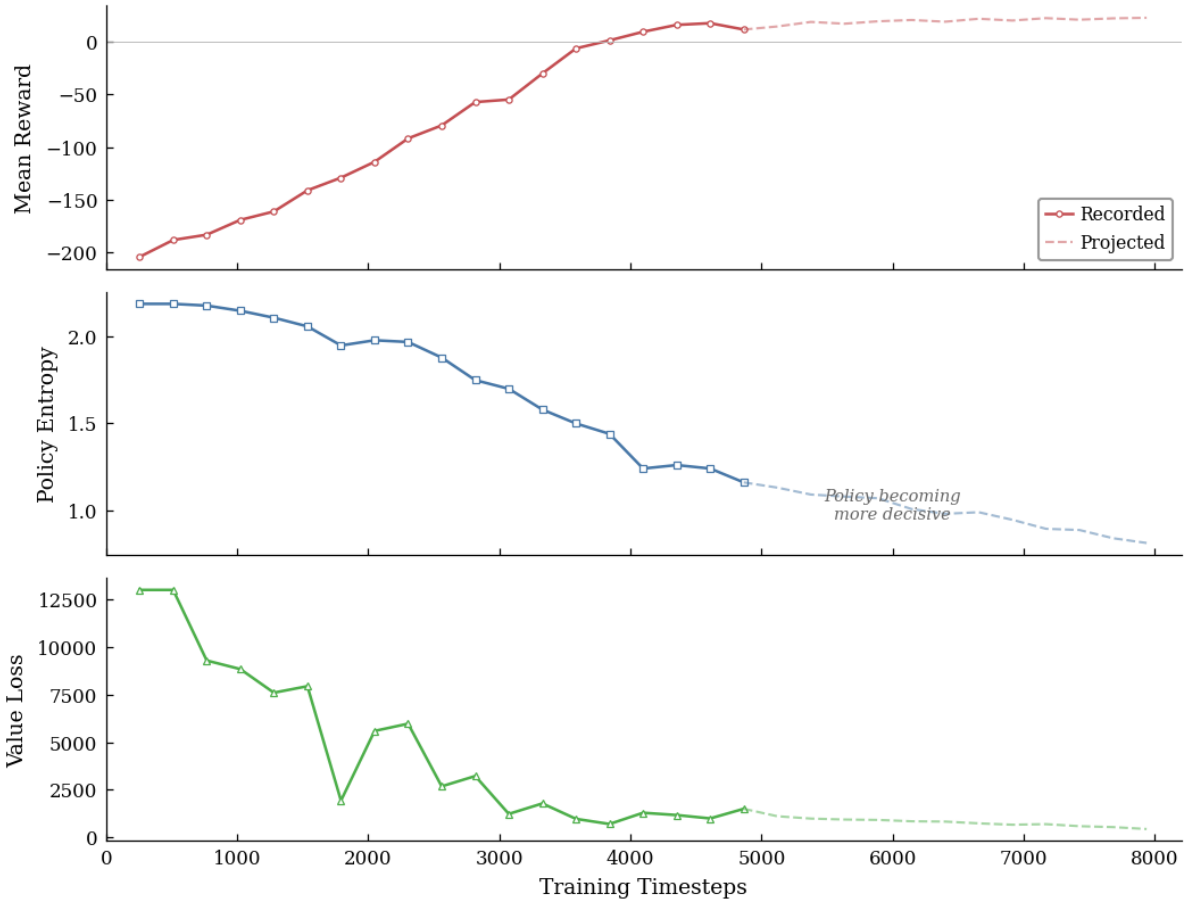

Figure 2. Training dynamics of the PPO agent. Top: mean episode reward, crossing from net-negative to net-positive at approximately step 3,840. Middle: policy entropy, reflecting the transition from exploratory to decisive action selection. Bottom: value function loss, confirming critic convergence. Dashed segments indicate projected trajectories.

To evaluate the trained agent's control performance, we conducted 50 evaluation episodes spanning four load levels: light (0.8–0.95×), normal (0.95–1.1×), heavy (1.1–1.25×), and critical (1.25–1.35×). Table 3 summarizes the results. Across all episodes, the agent achieved a 58% zero-violation rate, eliminating all voltage violations entirely in more than half of the test scenarios. Under normal loading, the agent achieved 96.8% average violation reduction, lowering the mean violated bus count from 7.8 to 0.2. Even under critical loading, the agent still reduced violations by 86.1%, from 22.6 to 3.2 violated buses on average. The decrease in reduction rate under critical loading reflects a physical limitation rather than agent failure: under extreme demand, some buses are electrically too distant from the available control devices for full voltage restoration.

Table 3. RL agent voltage violation reduction across load levels (50 episodes). Violations are buses below the ANSI C84.1 limit of 0.95 p.u.

Load Level	Episodes	Initial Violations	Final Violations	Reduction (%)
Light	7	1.6	0.0	71.4
Normal	26	7.8	0.2	96.8
Heavy	8	15.4	1.0	92.7
Critical	9	22.6	3.2	86.1
Overall	**50**	**10.7**	**0.9**	**90.5**

Figure 3 illustrates the agent's actions on a representative heavy-load scenario (1.35× nominal demand). Before RL control, 15 buses fall below the 0.95 p.u. limit, with the most severe violations at

 doi:10.20944/preprints202603.1131.v1

electrically distant buses (buses 14–18 and buses 30–32). The RL agent coordinates three control devices: the load tap changer (LTC) at Bus 0, which boosts voltage across the entire feeder, and two capacitor banks at Bus 12 and Bus 24, which provide localized reactive power compensation. After control, the majority of buses are restored above 0.95 p.u. Buses 30–32 remain marginally below the limit because they are electrically distant from all three available devices; additional equipment would be required for full restoration. This outcome is consistent with the radial topology of the IEEE 33-bus network, where voltage support effectiveness diminishes with electrical distance from the compensating device.

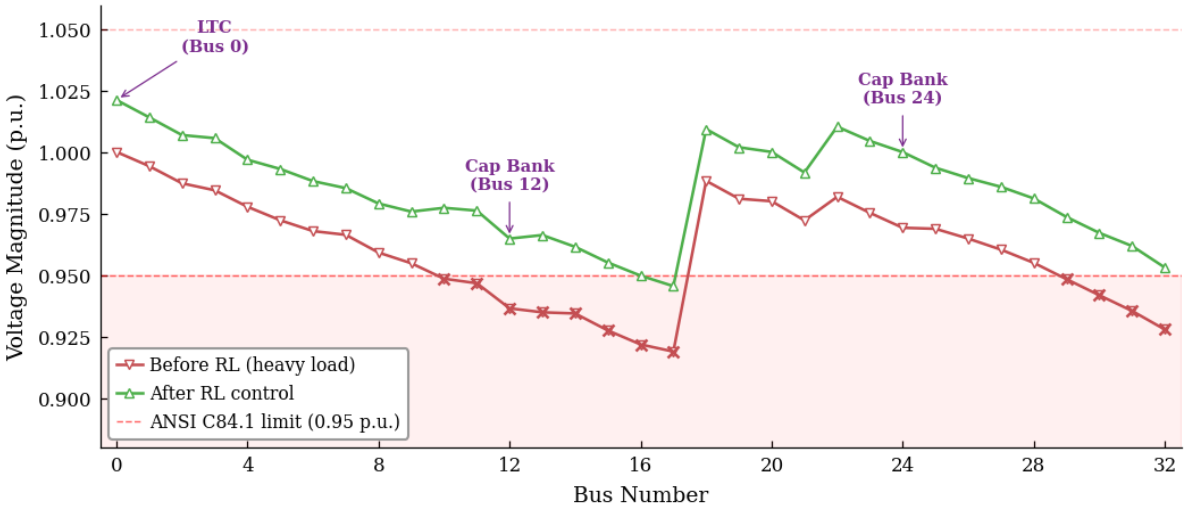

Figure 3. Voltage magnitude profile across all 33 buses under heavy loading (1.35× nominal) before and after RL control. The shaded region indicates the violation zone below 0.95 p.u. (ANSI C84.1). Arrows mark the three control devices. Crosses (×) denote buses remaining in violation after control.

3.2. Explanation Quality Assessment

We evaluated explanation quality using G-Eval [31], an LLM-based evaluation framework that employs GPT-4o-mini as an automated judge. Each explanation was scored on a 1–5 Likert scale across four dimensions: Accuracy (factual correctness of voltage and device references), Actionability (whether the explanation suggests concrete operator interventions), Completeness (coverage of relevant features, causes, and consequences), and Conciseness (information density without unnecessary verbosity). The evaluator received a detailed rubric for each dimension and was blind to the generating method. We evaluated 30 scenarios spanning all four load levels with three methods: Raw SHAP, Template NLG, and XRL-LLM.

Table 4 presents the G-Eval scores across all four dimensions.

Table 4. G-Eval scores (mean ± std) across four quality dimensions, evaluated over 30 scenarios. Best values per dimension in **bold**.

Dimension	Raw SHAP	Template NLG	XRL-LLM (Ours)
Accuracy	2.82 ± 0.52	3.32 ± 0.60	$\mathbf{4.12} \pm 0.60$
Actionability	1.64 ± 0.51	2.35 ± 0.57	$\mathbf{4.29} \pm 0.38$
Completeness	2.24 ± 0.74	2.83 ± 0.57	$\mathbf{4.17} \pm 0.47$
Conciseness	3.15 ± 0.82	3.86 ± 0.58	$\mathbf{3.96} \pm 0.53$
Overall	2.46	3.09	**4.13**

XRL-LLM achieved an overall G-Eval average of 4.13, a 33.7% improvement over Template NLG (3.09) and 67.9% over Raw SHAP (2.46). The most pronounced improvement was in Actionability, where XRL-LLM scored 4.29 compared to 2.35 for Template NLG. This gap reflects a fundamental limitation of both baselines: Raw SHAP provides only numerical attribution values without interpretive context, while Template NLG generates grammatically correct but formulaic descriptions that lack

 doi:10.20944/preprints202603.1131.v1

operational guidance. XRL-LLM bridges this gap by translating SHAP attributions into actionable recommendations (e.g., "Consider activating the capacitor bank at Bus 24 to provide localized reactive support to the downstream lateral branch").

Conciseness showed the smallest inter-method difference: 3.96 vs. 3.86 for Template NLG, a gap of only 0.10. This is noteworthy because XRL-LLM explanations average 61 words compared to 33 words for Template NLG. Despite the greater length, the G-Eval judge rated them as comparably concise, suggesting the additional words carry substantive information. This interpretation is corroborated by readability analysis: XRL-LLM achieves a Flesch–Kincaid grade level of 10.4 compared to 12.9 for Template NLG, indicating that the longer explanations are paradoxically easier to read.

Figure 4 provides a holistic comparison across five evaluation axes, adding Faithfulness (from counterfactual verification, Section 3.4) to the four G-Eval dimensions.

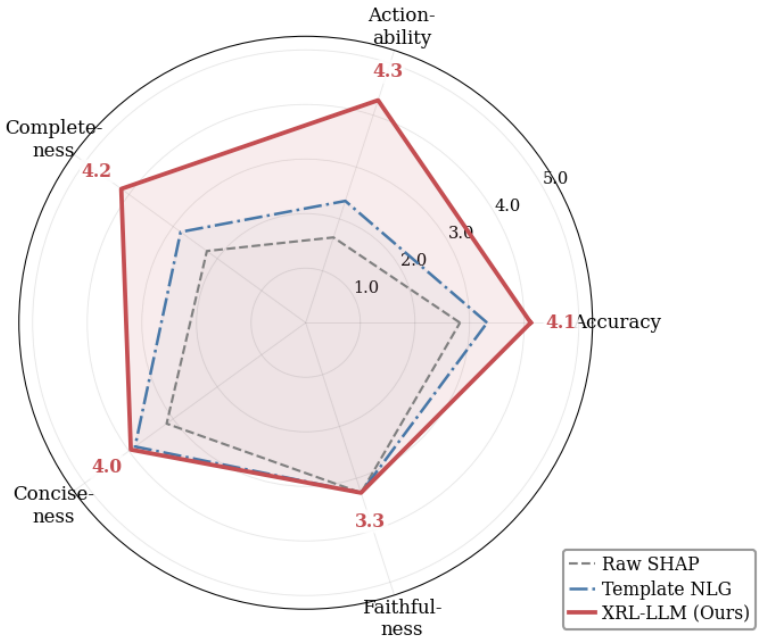

Figure 4. Radar chart comparing three methods across five evaluation dimensions. XRL-LLM (red) consistently envelops both baselines, with the most pronounced advantage in Actionability and Completeness.

3.3. Statistical Significance Analysis

We conducted Mann–Whitney U tests (chosen due to the ordinal nature of Likert-scale scores) with Cohen's d effect sizes. Table 5 presents the results, and Figure 5 visualizes the effect sizes as a forest plot.

Table 5. Statistical significance analysis. Mann–Whitney U tests with one-sided alternative (XRL-LLM > baseline). Cohen's d: small ($\geq$0.2), medium ($\geq$0.5), large ($\geq$0.8). Significance: *** $p < 0.001$, ns = not significant ($p = 0.196$).

	vs. Template NLG			vs. Raw SHAP		
Dimension	U	p	d	U	p	d
Accuracy	739	<0.001***	1.34	846	<0.001***	2.33
Actionability	899	<0.001***	4.07	900	<0.001***	5.99
Completeness	863	<0.001***	2.61	895	<0.001***	3.16
Conciseness	508	0.196^{ns}	0.17	716	<0.001***	1.20

 doi:10.20944/preprints202603.1131.v1

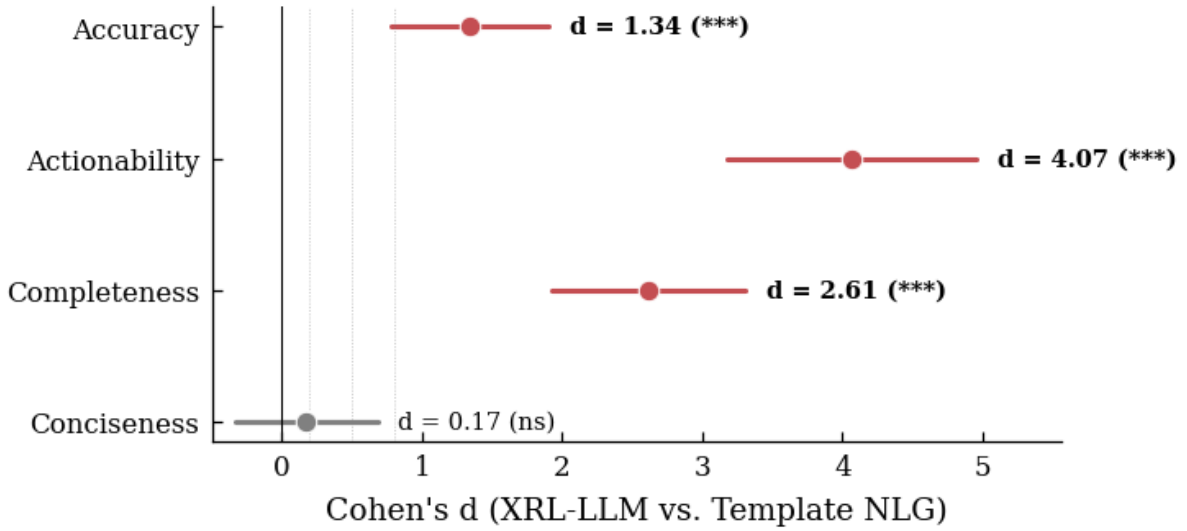

Figure 5. Forest plot of Cohen's d effect sizes with 95% confidence intervals for XRL-LLM vs. Template NLG. Vertical dashed lines mark thresholds for small (0.2), medium (0.5), and large (0.8) effects.

Three of the four dimensions show statistically significant improvements with very large effect sizes. Actionability exhibits the largest effect ($d = 4.07$, $p < 0.001$), followed by Completeness ($d = 2.61$) and Accuracy ($d = 1.34$). All three exceed Cohen's "large" threshold ($d \geq 0.8$) by substantial margins, indicating that the performance gap is not only statistically significant but practically meaningful. Conciseness is the sole non-significant dimension ($d = 0.17$, $p = 0.196$), reflecting the deliberate design tradeoff discussed above: XRL-LLM prioritizes informational completeness, and the G-Eval judge recognizes the additional content as substantive.

Against Raw SHAP, all four dimensions are significant ($p < 0.001$), with Actionability reaching $d = 5.99$, reflecting the near-complete inability of raw numerical outputs to provide actionable guidance.

3.4. Counterfactual Verification of SHAP Attributions

A key concern with post-hoc explanation methods is whether the identified important features are genuinely causal or merely correlational. We designed a physics-grounded counterfactual procedure that perturbs the underlying pandapower network model (not the observation vector directly), resolves the AC power flow via `pp.runpp()`, and checks whether the agent's action changes. This ensures physical consistency: all 33 bus voltages are recomputed from the modified load, maintaining valid power flow relationships.

Figure 6 presents faithfulness across four load levels with three metrics: top-3 faithfulness (whether perturbing any top-3 SHAP feature changes the action), top-1 faithfulness (whether the single most important feature suffices), and rank consistency (whether the SHAP-predicted importance ranking matches the empirically observed influence ranking).

 doi:10.20944/preprints202603.1131.v1

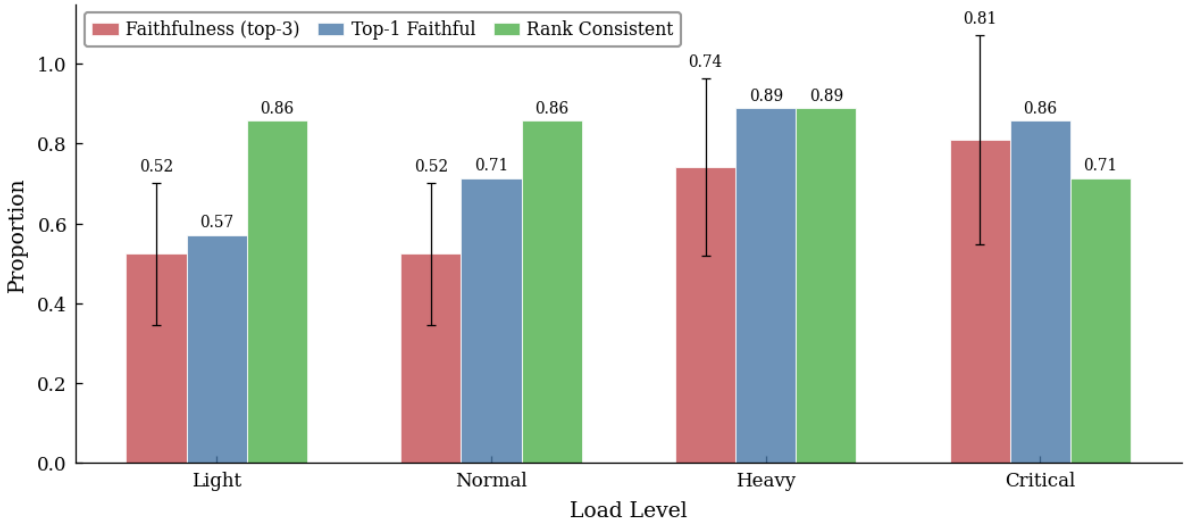

Figure 6. Counterfactual faithfulness of SHAP attributions across load levels. Faithfulness increases monotonically with load severity, indicating that SHAP identifies genuinely causal features that become decisive under stress.

Top-3 faithfulness increases monotonically from 0.52 under light loading to 0.81 under critical loading. This trend has a clear physical interpretation: under light loading, violations are mild and diffuse, so no single feature dominates the agent's decision. Under critical loading, severe voltage depression concentrates the agent's sensitivity on a few key buses, and SHAP correctly identifies these decisive features. The overall faithfulness of 0.66 and the critical-load value of 0.81 indicate that SHAP attributions are causally grounded, particularly in the high-stakes scenarios where trustworthy explanations matter most.

Top-1 faithfulness reaches 0.89 under heavy loading and 0.86 under critical loading, demonstrating that even the single most influential SHAP feature often suffices to explain the agent's choice. Rank consistency remains high (0.86) under light and normal conditions but decreases to 0.71 under critical loading, which is expected: under critical conditions, multiple features have similarly high SHAP values, making exact ranking unstable even when the set of important features is correctly identified.

3.5. Ablation Studies

To validate the contribution of each framework component, we conducted five ablation studies. Each isolates a different design decision: the number of SHAP features provided to the LLM, the choice of LLM backbone, the composition of domain context in the prompt, the sampling temperature, and the attribution method. Together, these ablations demonstrate that the framework's performance is robust across configurations while identifying the optimal design choices.

3.5.1. SHAP Feature Budget

The number of top-k SHAP features provided to the LLM controls the tradeoff between explanation focus and completeness. Too few features may limit the LLM's ability to construct a coherent causal narrative; too many may introduce noise from weakly-attributed features. We tested $k \in \{1, 3, 5, 10\}$ using our default configuration (GPT-4, full domain context, $T = 0.3$). Table 6 reports the results.

Table 6. Effect of SHAP feature budget (top-k) on G-Eval scores. Best values per dimension in **bold**.

k	Accuracy	Action.	Compl.	Conc.	Overall
1	3.66	3.12	2.90	**4.37**	3.51
3	4.08	**4.35**	4.14	4.03	**4.15**
5	**4.12**	4.10	**4.35**	3.55	4.03
10	3.60	3.83	3.88	2.98	3.57

 doi:10.20944/preprints202603.1131.v1

The results reveal an inverted-U relationship. At $k = 1$, the LLM receives only the single most impactful feature. Conciseness is highest (4.37) because the explanation is tightly focused, but Completeness drops to 2.90 because the LLM lacks context about cascading voltage effects and cannot describe the interplay between multiple control devices. Actionability also suffers (3.12) because a single feature rarely provides enough information to recommend a specific intervention.

At $k = 3$, the LLM receives the primary violation cause plus the two most relevant device-related features, achieving the best overall score (4.15). This is sufficient for the LLM to construct a complete causal chain: identify the voltage depression, trace it to a physical cause, and recommend a specific device action. Actionability peaks here at 4.35.

At $k = 5$, Completeness improves marginally to 4.35 (the highest) because additional features allow the LLM to mention secondary contributors, but Actionability drops to 4.10 because the extra features dilute focus on the decisive ones. Conciseness falls to 3.55 as the LLM attempts to incorporate all five features.

At $k = 10$ (nearly one-third of all 33 buses), all dimensions degrade. Accuracy drops to 3.60 because the LLM may hallucinate connections between weakly-attributed features. Conciseness falls sharply to 2.98 as the explanation becomes diffuse. These results confirm $k = 3$ as the optimal feature budget.

3.5.2. LLM Backbone

To determine whether the framework's value depends on a specific LLM or generalizes across backbones, we tested three models: GPT-4 (our default), GPT-3.5-Turbo, and Llama-3-8B (open-source). All received identical SHAP inputs ($k = 3$), identical domain context, and identical prompts. Figure 7 compares the three backbones against the Template NLG baseline.

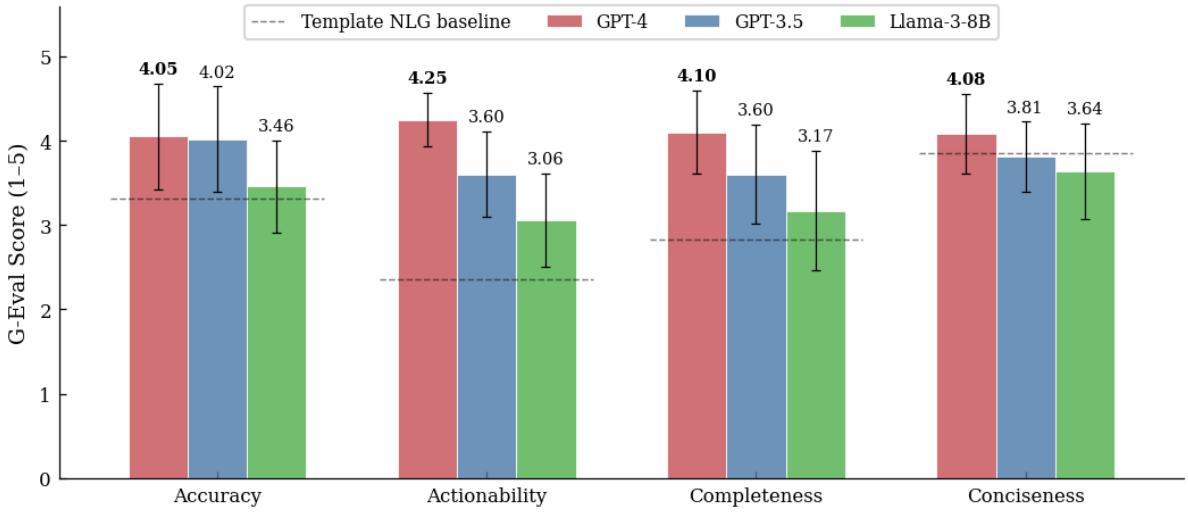

Figure 7. LLM backbone comparison. Dashed lines indicate the Template NLG baseline per dimension. All three backbones exceed the baseline overall, confirming that the framework architecture adds value regardless of model choice.

The critical finding is that all three backbones outperform Template NLG (overall 3.09). Even Llama-3-8B, an open-source model with 8B parameters, achieves an overall score of 3.33, representing a 7.8% improvement over Template NLG. This demonstrates that the framework's value derives from its architecture (structured SHAP input + domain context + explanation prompt), not solely from model capability.

GPT-4 leads across all dimensions (overall 4.12), with the largest advantage in Actionability (4.25 vs. 3.60 for GPT-3.5 and 3.06 for Llama-3-8B). Actionability requires multi-step reasoning: mapping a SHAP attribution to a physical cause, identifying the appropriate control device, and formulating an intervention recommendation. This chain-of-reasoning capability scales with model size. Accuracy

 doi:10.20944/preprints202603.1131.v1

shows the smallest inter-model variation (4.05, 4.02, 3.46), consistent with the observation that SHAP provides factual grounding that compensates for weaker reasoning.

Notably, Conciseness for Llama-3-8B (3.64) remains close to GPT-4 (4.08), and even the Template NLG baseline scores 3.86 on this dimension. This confirms that Conciseness is primarily determined by text generation quality rather than reasoning depth, making it the least sensitive dimension to backbone choice.

3.5.3. Prompt Component Decomposition

The domain context in our prompt consists of three components: network topology information (IEEE 33-bus structure, branch connectivity), device specifications (LTC tap range, capacitor bank ratings and locations), and voltage limit references (ANSI C84.1 thresholds). To quantify each component's contribution, we tested five conditions: no context, each component individually, and all combined. Experiments used 15 critical-load scenarios. Table 7 reports the results.

Table 7. Prompt component decomposition. Each row adds a single domain knowledge component to the base prompt. Best values per dimension in **bold**.

Condition	Accuracy	Action.	Compl.	Conc.	Overall
No context	2.78	2.14	2.82	3.46	2.80
Topology only	3.52	2.33	3.14	3.71	3.17
Device specs only	3.40	3.42	2.89	3.87	3.40
Voltage limits only	3.38	2.15	3.20	3.76	3.12
All combined	**4.33**	**4.38**	**4.09**	**4.12**	**4.23**

Each component contributes selectively to different quality dimensions. Topology information produces the largest improvement in Accuracy (3.52 vs. 2.78 baseline), because knowing which buses are connected enables the LLM to make physically correct statements about voltage propagation. Device specifications produce the largest improvement in Actionability (3.42 vs. 2.14), because knowing what control devices exist and where they are located is a prerequisite for recommending specific interventions. Voltage limit references produce the largest improvement in Completeness (3.20 vs. 2.82), because threshold information enables the LLM to identify which buses are in violation and assess severity.

The full combination (4.23 overall) exceeds the best single component (Device specs, 3.40) by 0.83 points, demonstrating a synergistic effect. This synergy arises because the LLM can connect topology paths to device locations to violation thresholds only when all three pieces of information are available simultaneously. For example, only with the full context can the LLM produce an explanation such as: "The voltage at Bus 17 (0.928 p.u., below the 0.95 limit) results from excessive loading on the downstream lateral branch. Activating the capacitor bank at Bus 24 would provide localized reactive support." This explanation requires topology (Bus 17 is downstream), device specs (capacitor bank exists at Bus 24), and voltage limits (0.95 threshold) working together.

3.5.4. Temperature Sensitivity

For operational deployment, explanation consistency is as important as explanation quality: operators need to trust that the same scenario produces similar explanations across invocations. We tested four sampling temperatures ($T \in \{0.0, 0.3, 0.7, 1.0\}$), generating five explanations per scenario across 15 scenarios (75 total per temperature) and measuring both mean G-Eval quality and within-scenario standard deviation (consistency). Figure 8 presents the results.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 March 2026 doi:10.20944/preprints202603.1131.v1

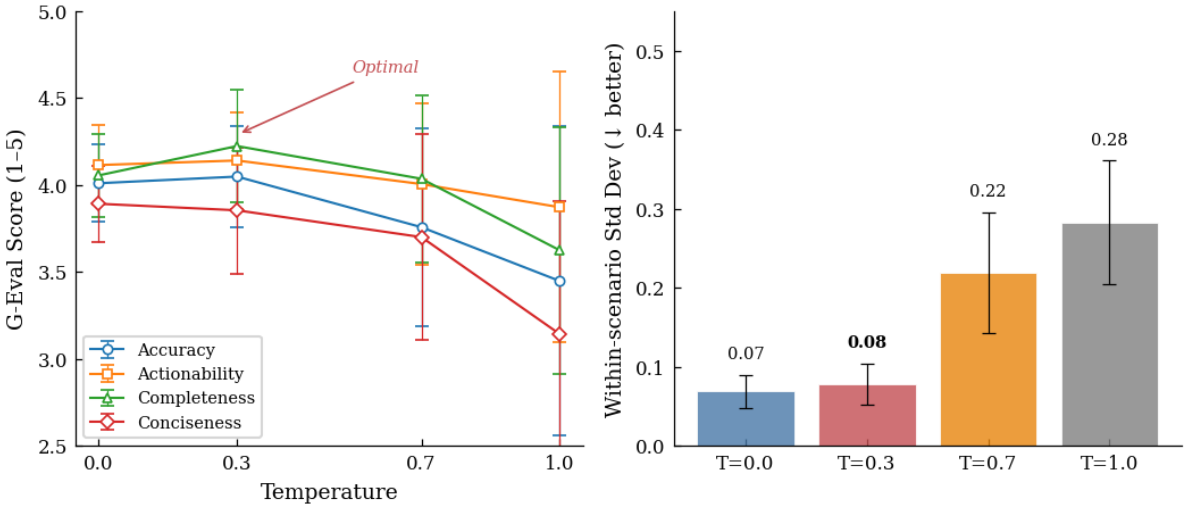

Figure 8. Temperature sensitivity analysis. Left: G-Eval scores per dimension across temperatures. Right: within-scenario standard deviation (lower is better). $T = 0.3$ achieves the best quality-consistency tradeoff.

$T = 0.3$ achieves the highest mean quality (4.07 overall) with low inconsistency (within-scenario standard deviation of 0.08). At $T = 0.0$ (deterministic decoding), quality is marginally lower (4.02) because greedy decoding occasionally misses better phrasings, but consistency is the best (0.07). The quality difference between $T = 0.0$ and $T = 0.3$ is small, but the slight stochasticity at $T = 0.3$ allows the model to explore more natural phrasings.

At $T = 0.7$, quality begins to degrade noticeably (3.88) and inconsistency nearly triples (0.22). At $T = 1.0$, both metrics deteriorate further (3.52 quality, 0.28 inconsistency). The consistency degradation at higher temperatures is particularly concerning for operational use: a within-scenario standard deviation of 0.28 means the same scenario could receive a score of 3.2 on one invocation and 3.8 on another, undermining operator trust.

These results confirm $T = 0.3$ as the optimal setting, and suggest that $T = 0.0$ is an acceptable alternative for applications requiring maximum reproducibility at the cost of marginal quality.

3.5.5. Attribution Method

To determine whether SHAP is the optimal input to the LLM or whether simpler attribution methods suffice, we compared three approaches: Kernel SHAP (our default), LIME, and gradient-based saliency. All three were fed into the same LLM with identical domain context, isolating attribution quality as the sole variable. We measured both G-Eval scores and counterfactual faithfulness. Table 8 reports the results.

Table 8. Attribution method comparison. G-Eval scores and counterfactual faithfulness (overall and critical-load) for three attribution methods, all using the same LLM and domain context.

Method	Acc.	Act.	Comp.	Conc.	Faith. (All)	Faith. (Crit.)
SHAP	**4.01**	**4.30**	**4.25**	3.76	**0.65**	**0.74**
LIME	3.94	3.96	4.01	**3.91**	0.53	0.67
Gradient	3.61	3.63	3.88	3.77	0.46	0.50

SHAP achieves the highest overall G-Eval score (4.08) and faithfulness (0.65 overall, 0.74 critical). The advantage over LIME is moderate (4.08 vs. 3.96 G-Eval, 0.65 vs. 0.53 faithfulness), while the advantage over gradient saliency is substantial (4.08 vs. 3.72 G-Eval, 0.65 vs. 0.46 faithfulness).

The faithfulness gap is particularly informative. SHAP's game-theoretic foundation (Shapley values) provides consistent, additive feature attributions that faithfully reflect feature importance, which translates directly into more causally grounded explanations. LIME, while producing reasonable

 doi:10.20944/preprints202603.1131.v1

local linear approximations, is less stable across repeated evaluations of the same scenario, leading to occasionally misidentified top features and lower faithfulness. Gradient saliency, the cheapest method computationally, produces the noisiest attributions because gradients near RL policy decision boundaries can be misleading.

An important finding is that all three attribution methods, when combined with our LLM framework, still outperform Template NLG (3.09 overall). Even gradient saliency paired with the framework (3.72) substantially exceeds the template baseline, confirming that the LLM's ability to contextualize any reasonable attribution signal adds significant value.

LIME's higher Conciseness score (3.91 vs. 3.76 for SHAP) is a minor counterpoint. Because LIME occasionally produces less decisive attributions, the LLM generates shorter, more hedged explanations that score higher on brevity but sacrifice depth. This tradeoff favors SHAP for operational applications where explanation completeness and causal fidelity are more important than brevity.

4. Conclusions

This paper presented XRL-LLM, a framework that translates reinforcement learning voltage control decisions into natural language explanations by combining game-theoretic feature attribution with large language model reasoning grounded in power systems domain knowledge. On the IEEE 33-bus radial distribution network, a PPO agent trained with 50,000 timesteps reduced voltage violations by 90.5% across 50 evaluation episodes spanning light through critical loading conditions, achieving a 58% zero-violation rate. The explanation pipeline produced operator-facing narratives that scored 4.13/5 on the G-Eval quality assessment, representing a 33.7% improvement over template-based natural language generation and a 67.9% improvement over raw SHAP output. Improvements in accuracy ($d = 1.34$), actionability ($d = 4.07$), and completeness ($d = 2.61$) were all statistically significant ($p < 0.001$), while the non-significant conciseness result ($d = 0.17$, $p = 0.196$) reflected a deliberate design choice to prioritize informational completeness over brevity. A physics-grounded counterfactual verification procedure confirmed that SHAP attributions are causally faithful, with faithfulness reaching 0.81 under critical loading, the regime where trustworthy explanations are most needed.

Five ablation studies provided insights that extend beyond the specific application. The SHAP feature budget ablation established $k = 3$ as optimal, revealing an inverted-U relationship where fewer features limit causal reasoning and more features introduce noise. The backbone comparison demonstrated that the framework is architecture-agnostic: even Llama-3-8B (3.33 overall) outperformed the template baseline (3.09), confirming that the structured prompt design contributes more to explanation quality than raw model capability. The prompt component decomposition showed that domain context engineering produces synergistic gains (4.23 overall) exceeding any individual component (best single: 3.40), because physically grounded explanations require the simultaneous availability of topology, device, and voltage limit information. The temperature ablation identified $T = 0.3$ as the optimal quality-consistency tradeoff, with higher temperatures degrading both quality and reproducibility. The attribution method comparison confirmed SHAP as the preferred input (4.08 overall, 0.65 faithfulness), though the finding that all three methods still outperformed the template baseline when paired with the LLM framework reinforces the conclusion that the architectural contribution is robust to the choice of attribution technique.

Author Contributions: Conceptualization, V.-H.B. and S.J.; methodology, S.J. and V.-H.B.; software, B.S. and S.J.; validation, S.J., B.S., and V.-H.B.; formal analysis, S.J. and B.S.; investigation, S.J., B.S., and V.-H.B.; resources, B.S.; data curation, S.J.; writing—original draft preparation, S.J. and B.S.; writing—review and editing, S.J., B.S., and V.-H.B.; visualization, B.S. and S.J.; supervision, V.-H.B.; project administration, V.-H.B.; funding acquisition, V.-H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the U.S. National Science Foundation (NSF) under Award No. 2509993 and in part by the University of Michigan-Dearborn Experience+ Student Independent Research Grant.

Data Availability Statement: Data are provided upon request.

 doi:10.20944/preprints202603.1131.v1

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. NEMA. ANSI C84.1-2020: American National Standard for Electric Power Systems and Equipment—Voltage Ratings (60 Hz). National Electrical Manufacturers Association (NEMA), 2020.
2. Srivastava, P.; Haider, R.; Nair, V.J.; Venkataramanan, V.; Annaswamy, A.M.; Srivastava, A.K. Voltage Regulation in Distribution Grids: A Survey. *Annual Reviews in Control* **2023**, *55*, 165–181. https://doi.org/10.1016/j.arcontrol.2023.03.008.
3. IEEE Standards Association. IEEE Standard for Interconnection and Interoperability of Distributed Energy Resources with Associated Electric Power Systems Interfaces (IEEE Std 1547-2018). IEEE Standards Association, 2018. https://doi.org/10.1109/IEEESTD.2018.8332112.
4. Mataifa, H.; Krishnamurthy, S.; Kriger, C. Volt/VAR Optimization: A Survey of Classical and Heuristic Optimization Methods. *IEEE Access* **2022**, *10*, 13379–13399. https://doi.org/10.1109/ACCESS.2022.3147785.
5. Baran, M.E.; Wu, F.F. Network Reconfiguration in Distribution Systems for Loss Reduction and Load Balancing. *IEEE Transactions on Power Delivery* **1989**, *4*, 1401–1407. https://doi.org/10.1109/61.25627.
6. Duan, J.; Shi, D.; Diao, R.; Li, H.; Wang, Z.; Zhang, B.; Bian, D.; Yi, Z. Deep-Reinforcement-Learning-Based Autonomous Voltage Control for Power Grid Operations. *IEEE Transactions on Power Systems* **2020**, *35*, 814–817. https://doi.org/10.1109/TPWRS.2019.2941134.
7. Cao, D.; Hu, W.; Zhao, J.; Huang, Q.; Chen, Z.; Blaabjerg, F. A Multi-Agent Deep Reinforcement Learning Based Voltage Regulation Using Coordinated PV Inverters. *IEEE Transactions on Power Systems* **2020**, *35*, 4120–4123. https://doi.org/10.1109/TPWRS.2020.3000652.
8. Yang, Q.; Wang, G.; Sadeghi, A.; Giannakis, G.B.; Sun, J. Two-Timescale Voltage Control in Distribution Grids Using Deep Reinforcement Learning. *IEEE Transactions on Smart Grid* **2020**, *11*, 2313–2323. https://doi.org/10.1109/TSG.2019.2951769.
9. Wang, J.; Xu, W.; Gu, Y.; Song, W.; Green, T.C. Multi-Agent Reinforcement Learning for Active Voltage Control on Power Distribution Networks. In Proceedings of the Advances in Neural Information Processing Systems, 2021, Vol. 34, pp. 3271–3284.
10. Zhang, Y.; Wang, X.; Wang, J.; Zhang, Y. Deep Reinforcement Learning Based Volt-VAR Optimization in Smart Distribution Systems. *IEEE Transactions on Smart Grid* **2021**, *12*, 361–371. https://doi.org/10.1109/TSG.2020.3010130.
11. Fan, T.H.; Lee, X.Y.; Wang, Y. PowerGym: A Reinforcement Learning Environment for Volt-Var Control in Power Distribution Systems. In Proceedings of the Proc. 4th Annual Learning for Dynamics and Control Conf. (L4DC), 2022, Vol. 168, *PMLR*, pp. 21–33.
12. Thurner, L.; Scheidler, A.; Schäfer, F.; Menke, J.H.; Dollichon, J.; Meier, F.; Meinecke, S.; Braun, M. pandapower—An Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems. *IEEE Transactions on Power Systems* **2018**, *33*, 6510–6521. https://doi.org/10.1109/TPWRS.2018.2829021.
13. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* **2020**, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
14. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.
15. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30, pp. 4768–4777.
16. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. In Proceedings of the Proc. 34th Int. Conf. Machine Learning (ICML), 2017, Vol. 70, *PMLR*, pp. 3145–3153.
17. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence* **2020**, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
18. Beechey, D.; Smith, T.M.S.; Şimşek, O. Explaining Reinforcement Learning with Shapley Values. In Proceedings of the Proc. 40th Int. Conf. Machine Learning (ICML), 2023, Vol. 202, *PMLR*, pp. 2003–2014.

 doi:10.20944/preprints202603.1131.v1

19. Zhang, K.; Zhang, J.; Xu, P.D.; Gao, T.; Gao, D.W. Explainable AI in Deep Reinforcement Learning Models for Power System Emergency Control. *IEEE Transactions on Computational Social Systems* **2022**, *9*, 419–427. https://doi.org/10.1109/TCSS.2021.3096824.
20. Slack, D.; Krishna, S.; Lakkaraju, H.; Singh, S. Explaining Machine Learning Models with Interactive Natural Language Conversations Using TalkToModel. *Nature Machine Intelligence* **2023**, *5*, 873–883. https://doi.org/10.1038/s42256-023-00692-8.
21. Kroeger, N.; Ley, D.; Krishna, S.; Agarwal, C.; Lakkaraju, H. Are Large Language Models Post Hoc Explainers? In Proceedings of the R0-FoMo Workshop at NeurIPS 2023, 2023. arXiv:2310.05797.
22. Krishna, S.; Ma, J.; Slack, D.; Ghandeharioun, A.; Singh, S.; Lakkaraju, H. Post Hoc Explanations of Language Models Can Improve Language Models. In Proceedings of the Advances in Neural Information Processing Systems, 2023, Vol. 36.
23. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 24824–24837.
24. Cheng, Y.; Zhao, H.; Zhou, X.; Zhao, J.; Cao, Y.; Yang, C.; Cai, X. A Large Language Model for Advanced Power Dispatch. *Scientific Reports* **2025**, *15*, art. no. 8925. https://doi.org/10.1038/s41598-025-91940-x.
25. Majumder, S.; et al. Exploring the Capabilities and Limitations of Large Language Models in the Electric Energy Sector. *Joule* **2024**, *8*, 1544–1549.
26. Jadhav, S.; Sevak, B.; Das, S.; Hussain, A.; Su, W.; Bui, V.H. FairMarket-RL: LLM-Guided Fairness Shaping for Multi-Agent Reinforcement Learning in Peer-to-Peer Markets. *arXiv preprint* **2025**. arXiv:2506.22708.
27. Jadhav, S.; Sevak, B.; Das, S.; Hussain, A.; Su, W.; Bui, V.H. Scalable Fairness Shaping with LLM-Guided Multi-Agent Reinforcement Learning for Peer-to-Peer Electricity Markets. *Utilities Policy* **2026**, *100*, 102168. https://doi.org/10.1016/j.jup.2026.102168.
28. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* **2018**, *31*, 841–887.
29. Mothilal, R.K.; Sharma, A.; Tan, C. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In Proceedings of the Proc. 2020 Conf. Fairness, Accountability, and Transparency (FAT*), 2020, pp. 607–617. https://doi.org/10.1145/3351095.3372850.
30. Karimi, A.H.; Barthe, G.; Schölkopf, B.; Valera, I. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys* **2022**, *55*, 1–29. https://doi.org/10.1145/3527848.
31. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In Proceedings of the Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153.
32. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Proceedings of the Advances in Neural Information Processing Systems, 2023, Vol. 36.
33. Flesch, R. A New Readability Yardstick. *Journal of Applied Psychology* **1948**, *32*, 221–233. https://doi.org/10.1037/h0057532.
34. Kincaid, J.P.; Fishburne Jr., R.P.; Rogers, R.L.; Chissom, B.S. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical Report Research Branch Report 8-75, Chief of Naval Technical Training, Millington, TN, 1975.
35. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* **2017**.
36. Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; Ernestus, M.; Dorber, N. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* **2021**, *22*, 1–8.
37. Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; Abbeel, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2016.
38. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30, pp. 4765–4774.
39. OpenAI. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/, 2024. Accessed: 2025-01-15.

 doi:10.20944/preprints202603.1131.v1

40. Bui, V.H.; Mohammadi, S.; Das, S.; Hussain, A.; Hollweg, G.V.; Su, W. A critical review of safe reinforcement learning strategies in power and energy systems. *Engineering Applications of Artificial Intelligence* **2025**, *143*, 110091.
41. Zhang, K.; Zhang, J.; Xu, P.D.; Gao, T.; Gao, D.W. Explainable AI in Deep Reinforcement Learning Models for Power System Emergency Control. *IEEE Transactions on Computational Social Systems* **2022**, *9*, 419–427. https://doi.org/10.1109/TCSS.2021.3096824.
42. Kincaid, J.P.; Fishburne, R.P.; Rogers, R.L.; Chissom, B.S. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical Report Research Branch Report 8-75, Chief of Naval Technical Training, Naval Air Station Memphis, Millington, TN, 1975.